SEGMENTAZIONE SEMI-AUTOMATICA DEL PARLATO MEDIANTE APPLICAZIONE DI UN MODELLO DEL SISTEMA UDITIVO PERIFERICO

Piero Cosi
Centro di Studio per le Ricerche di Fonetica C.N.R. Padova
P.zza Salvemini 13, 35131 - Padova (Italy)

SOMMARIO

Viene descritto un sistema semi-automatico di allineamento temporale del segnale vocale con la sua corrispondente trascrizione fonetica. Il sistema fornisce in modo automatico alcune ipotesi di segmentazione allo scopo di rendere più veloce il compito di esperti fonetisti nell'analizzare grosse basi dati. Sulla base della conoscenza ortografica del testo pronunciato gli esperti devono scegliere l'allineamento più opportuno fra quelli proposti automaticamente. Il sistema, che riceve in ingresso i parametri forniti da un modello del sistema uditivo periferico dimostratosi molto efficace nel codificare le informazioni contenute nel segnale vocale, si basa interamente sulla teoria della segmentazione multi-livello per la costruzione delle ipotesi di segmentazione.

Dopo una breve descrizione della procedura e dell'ambiente di acquisizione del segnale vocale di riferimento su cui sono basati tutti i lavori presentati in questo volume[1], viene schematicamente illustrato il modello del sistema uditivo periferico utilizzato quale tecnica di analisi in questo lavoro e, successivamente, viene descritto il sistema semi-automatico di segmentazione, realizzato sulla base di questo front-end uditivo, assieme anche ad alcuni dei risultati ottenuti.

INTRODUZIONE

L'allineamento temporale del segnale vocale con la sua corrispondente trascrizione fonetica è normalmente affidato all'opera manuale di esperti fonetisti. Nonostante l'ausilio di sempre più affidabili strumenti audio visivi, le divergenze nella segmentazione manuale dello stesso materiale vocale, effettuata da parte di più esperti, non potranno mai essere completamente eliminate. A causa delle diverse capacità percettive, sia visive che uditive, come anche dell'oggettiva difficoltà di definire una inequivocabile strategia comune, è evidente l'implicita incoerenza di un tale approccio manuale. Un altro svantaggio è dato dall'inevitabile spreco di risorse, sia temporali che umane. Sulla base di queste considerazioni, l'interesse per la realizzazione di sistemi automatici di segmentazione e "labelling" sta sempre più crescendo. Tali sistemi automatici, oltre a minimizzare i tempi di esecuzione, rendono implicitamente coerenti i risultati della segmentazione. Infatti, gli errori di segmentazione risultano facilmente identificabili e categorizzabili a causa della natura algoritmica delle procedure.

Il sistema descritto in questo lavoro fornisce in modo automatico alcune ipotesi di segmentazione allo scopo di ridurre al minimo il compito di esperti fonetisti nell'analizzare grosse basi dati. Nessun istante di segmentazione viene posizionato manualmente e agli esperti viene esclusivamente richiesta un'azione di supervisione sulle ipotesi di segmentazione prodotte automaticamente dal sistema. Gli esperti devono infatti scegliere, sulla base della conoscenza ortografica del testo pronunciato, l'allineamento più opportuno fra quelli proposti automaticamente, eventualmente eliminando "marker" sovrabbondanti.

DESCRIZIONE DELLA PROCEDURA E DEL SISTEMA DI ACQUISIZIONE

Il segnale vocale utilizzato come riferimento per i lavori presentati in questo volume è stato scelto sulla base di tre semplici criteri: naturalezza, durata e qualità. Il parlato in esame doveva essere infatti:

[1] Questo lavoro, assieme a tutti gli altri presentati in questo volume, è stato realizzato in occasione del Workshop del Gruppo di Fonetica Sperimentale, dedicato ad un confronto fra le varie tecniche di analisi del parlato, svoltosi ad Acavacata di Rende presso il Laboratorio di Fonetica dell'Università della Calabria, il 28-29 novembre 1991.

spontaneo, cioè non letto, non interrotto frequentemente da altri parlanti, di durata non inferiore ai 30 secondi (per consentire non solo analisi a livello acustico, ma anche a livello "linguistico"), e di buona qualità relativamente alle caratteristiche di rapporto segnale/disturbo.
Dopo varie sedute di ascolto campione è stata scelta una conversazione radiofonica della durata di circa 30 secondi, estratta dalla tramissione Orione del settembre 1991, trasmessa dal terzo canale Radio relativa ad una conversazione con l'onorevole Luigi Covatta, sottosegretario al Ministero dei Beni Culturali, avente come argomento un tema di attualità.

Come descritto in Fig. 1, il segnale è stato acquisito, tramite un sintonizzatore radio (Scott Stereo Tuner T 526L) direttamente su computer (Compaq 386/25) mediante una scheda di acquisizione (OROS AU21). La frequenza di acquisizione è stata di 20 kHz e la precisione di 16 bit, in quanto questi due valori sono diventati ormai lo standard europeo. Una versione per Macintosh è stata ottenuta convertendo il formato del segnale via software tramite il programma Audiomedia.

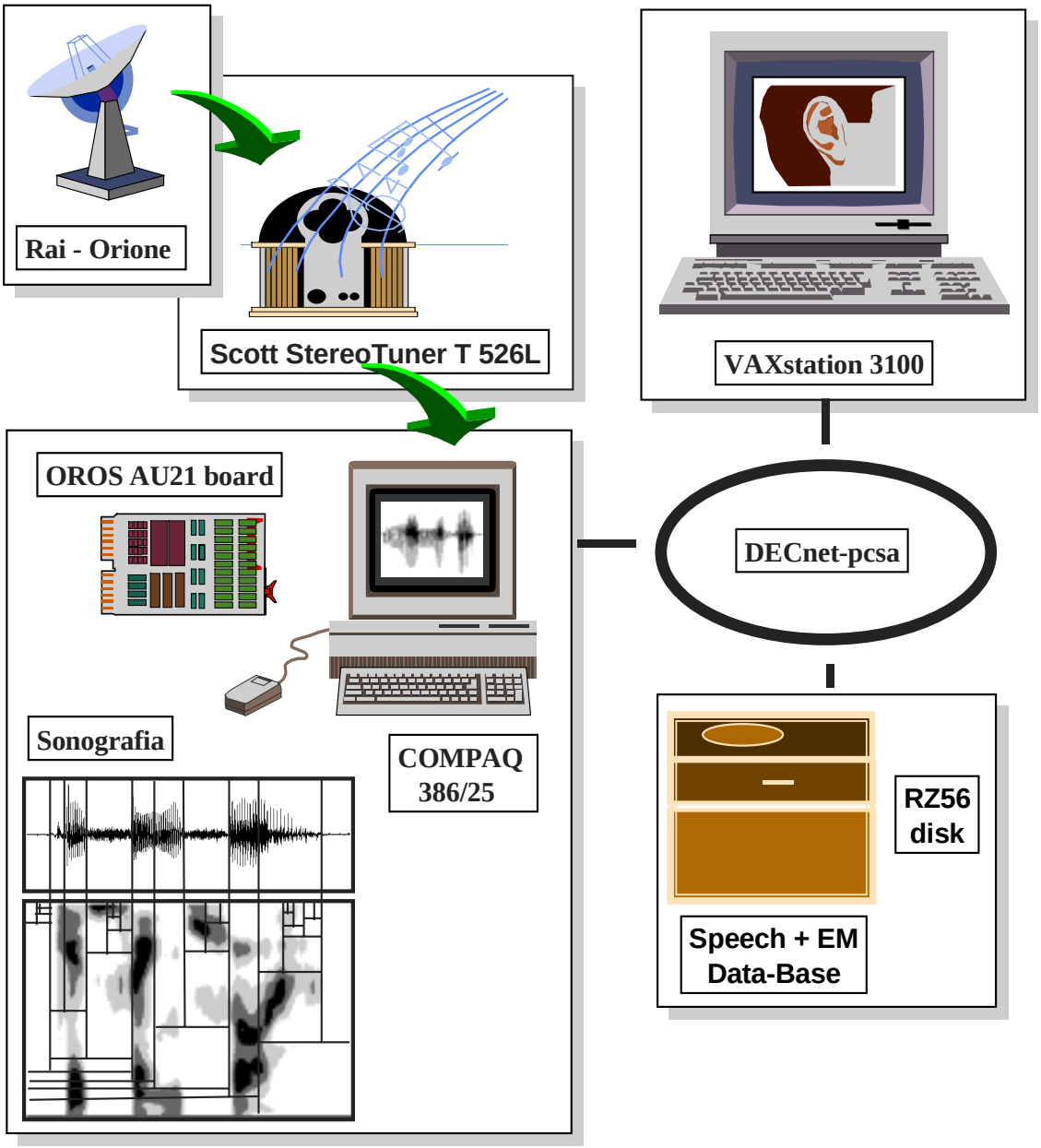

Figura 1. Diagramma a blocchi del sistema di acquisizione e segmentazione del parlato.

ARCHITETTURA DEL SISTEMA DI SEGMENTAZIONE

Il sistema di segmentazione utilizzato in questo lavoro si divide essenzialmente in tre parti. Ad una prima fase di elaborazione digitale del segnale vocale, eseguita mediante un Modello del Sistema Uditivo Periferico (**MSUP**) [1], dimostratosi estremamente efficace in problemi di segmentazione e classificazione fonetica [2],[3], segue una fase di individuazione sul segnale di vari possibili confini di separazione fra le varie unità e, sulla base di queste informazioni, nell'ultima fase viene richiesto ad un esperto di scegliere l'ipotesi più opportuna fra quelle proposte.

MODELLO DEL SISTEMA UDITIVO PERIFERICO (MSUP)

Il segnale vocale viene preelaborato tramite un MSUP essenzialmente analogo a quello sviluppato recentemente al Massachussetts Institute of Technology (MIT) da S. Seneff [1]. Il modello è implementato in FORTRAN su un minielaboratore (DEC VAXstation 3100) connesso in rete (DECNET-PCSA) ad un Personal Computer (Compaq 386/25) dedicato all'acquisizione del segnale vocale (vedi Fig. 1). Strutturalmente il modello, il cui schema a blocchi è illustrato in Fig. 2, è separato nella cascata di tre distinte sezioni di elaborazione: un banco di filtri a banda critica per l'analisi in frequenza, un modello dei meccanismi cellulari (sinapsi), e due moduli chiamati rivelatori di inviluppo e sincronia.

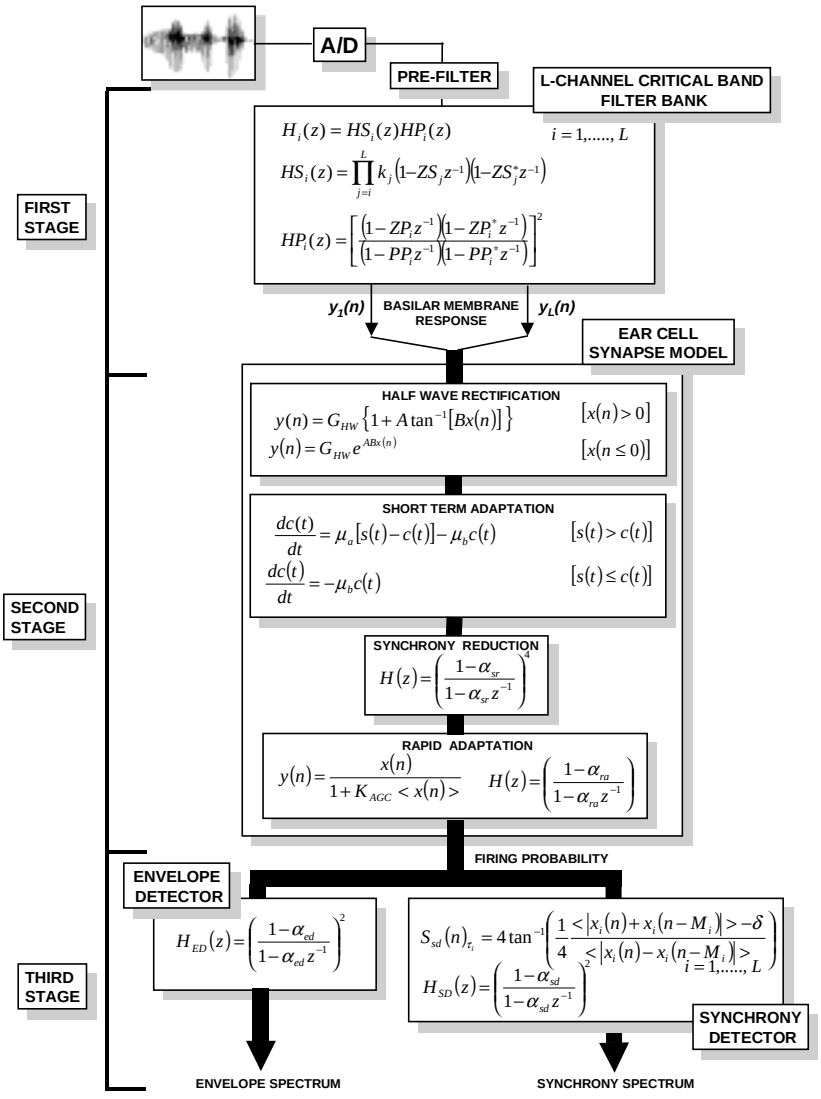

Figura 2. Descrizione del Modello del Sistema Uditivo Periferico.

Le prime due sezioni costituiscono la simulazione vera e propria dei processi che avvengono nel sistema uditivo dal punto della recezione del suono, fino al punto della sua codifica in scariche elettriche sull'VIII nervo acustico. La terza sezione non è il modello reale di alcun corrispettivo fisiologico, ma ha il compito di elaborare l'informazione contenuta nei pattern di scarica in modo da rendere più evidenti alcune grandezze di effettivo interesse percettivo. In Fig.2 è mostrato lo schema a blocchi dell'architettura globale del modello. L'implementazione di ogni modulo è rappresentata dalle formule matematiche indicate all'interno dei singoli blocchi di Fig. 2.

Nella **prima sezione** si opera in parallelo nel dominio della frequenza attraverso un **banco di 40 filtri a banda critica** che coprono una banda di frequenze che va circa da 100 Hz a 7500 Hz. In tale sezione viene simulata l'azione di filtraggio operata dalla membrana basilare sulle onde di pressione prodotte nella coclea. Il segnale d'ingresso, preventivamente campionato alla frequenza di 16 kHz[2], viene pre-filtrato per eliminare le frequenze superiori a 8 kHz. Successivamente viene inviato al banco di 40 filtri passa-banda lineari e tempo-invarianti, spaziati in frequenza in scala Bark, le cui risposte in frequenza riproducono le curve di sintonizzazione fisiologica delle fibre nervose coinvolte nella percezione acustica.

Nella **seconda sezione** del sistema, il **modello delle cellule cigliari e delle sinapsi**, vengono simulati i fenomeni che avvengono a livello elettro-neurale nel meccanismo di "trasduzione" delle vibrazioni meccaniche in scariche elettriche sulle fibre nervose. Questa seconda sezione, che si ripete in parallelo per ognuno dei 40 canali del modello, si può a sua volta scindere in 4 sotto-moduli posti in cascata, ognuno dei quali simula una particolare caratteristica del comportamento neurale:

a) modulo di raddrizzamento ad una semionda e di saturazione: simula la risposta delle cellule cigliari alle vibrazioni meccaniche attraverso la scarica di potenziali in un solo senso (raddrizzamento ad una semionda), ed inoltre simula la compressione dinamica del dominio di variazione (saturazione);

b) modulo di adattamento alla risposta: ha lo scopo di simulare il fenomeno di adattamento della risposta, cioè l'abbassamento del tasso di scarica nelle fibre nervose quando l'ingresso è uno stimolo sufficentemente prolungato; questo fenomeno si verifica a livello sinaptico tra le cellule cigliate e le fibre nervose afferenti;

c) modulo di riduzione della sincronizzazione: simula la perdita di sincronizzazione nelle sequenze di scarica quando lo stimolo supera una certa frequenza; questa sorta di "inerzia" nel seguire segnali veloci è imputabile al tempo minimo di latenza tra una scarica e l'altra in ogni singolo neurone; questo modulo è lineare ed è implementato tramite un filtro passa-basso;

d) modulo di controllo automatico del guadagno (CAG): ha lo scopo di simulare il fenomeno di adattamento rapido e cioè la rapida riduzione del tasso di scarica nella parte iniziale della risposta agli stimoli d'ingresso; questo effetto deriva dalla proprietà di refrattarietà delle fibre nervose.

Si noti la particolare struttura e l'ordine con cui si susseguono i vari moduli in questa seconda sezione del modello. Ciò rispecchia abbastanza fedelmente la struttura dei meccanismi fisiologici e la sequenza di trasformazioni a cui è soggetto lo stimolo acustico. Le uscite di questa seconda sezione del modello descrivono come varia la probabilità di scarica in gruppi di fibre sintonizzate alla frequenza centrale (CF) del relativo canale al variare del segnale d'ingresso.

Per ognuno dei canali si procede, nella **terza sezione**, a due elaborazioni in parallelo:

a) modulo rivelatore di inviluppo: opera un filtraggio di tipo passa-basso sul segnale d'uscita al secondo stadio in modo da fornirne l'inviluppo; questo corrisponde praticamente alla rilevazione del tasso medio di scarica nella fibra corrispondente al variare del tempo; l'andamento dell'inviluppo dipende dall'energia delle componenti armoniche presenti nella banda del canale;

b) modulo rivelatore di sincronia: ha la funzione di operare un'analisi della periodicità sul segnale d'uscita al secondo stadio; questo blocco, è in grado di rilevare la presenza di frequenze dominanti ("formanti") a frequenze centrate sul filtro passa-banda iniziale del relativo canale.

In Fig. 3, è illustrato un esempio dell'uscita del MSUP, per quanto riguarda il modulo rivelatore di inviluppo (b) e di sincronia (c), applicato alla frase inglese "...Susan ca(n'tgo)..." (a) pronunciata da un parlante femminile (vedi [4]). L'utilizzazione dei parametri relativi al modulo rilevatore di sincronia (Fig.

[2] Il segnale vocale utilizzato come riferimento in tutti i lavori di questo Volume è stato campionato a 20 kHz. Per utilizzare il MSUP quale front-end di analisi acustica in questo lavoro si è resa quindi necessaria un'opportuna conversione di frequenza da 16 a 20 kHz, che è stata effettuata via software.

3c) consentono di produrre degli "spettri" contenenti un numero limitato di linee spettrali ben definite, indicando una buona utilizzazione delle conoscenze sulla produzione e percezione del segnale vocale, in accordo alle quali, le "formanti" sono parametri estremamente stabili nel caratterizzare le vocali o in generale gli stimoli sonoranti.

Questo modello si è rivelato di estremo interesse anche nell'analizzare il segnale vocale in condizioni rumorose [5]. Fig. 4 si riferisce infatti all'analisi della stessa frase considerata precedentemente in Fig.3 a cui è stato sovrapposto un elevato livello di rumore [4]. Nonostante la notevole degradazione (Fig. 4a), la struttura formantica dello stimolo risulta essere ben preservata dall'analisi con il MSUP (Fig. 4c).

SISTEMA DI SEGMENTAZIONE MULTI-LIVELLO

Le uscite del MSUP sono state utilizzate in ingresso al vero e proprio algoritmo di segmentazione che si basa interamente sulla teoria della segmentazione multi-livello [6]. La filosofia alla base di questa teoria sottolinea che non esite un unico livello di rappresentazione segmentale in grado di descrivere tutti gli eventi acustici di interesse presenti nel segnale vocale. Per ovviare a questa implicita difficoltà viene adottata una rappresentazione multi-livello la quale consente di evidenziare all'interno di un unica struttura sia i mutamenti rapidi che quelli graduali riscontrabili sul segnale. La costruzione della struttura multi-livello [6] può riassumersi in due fasi. Nella prima fase vengono ricercati gli eventi acustici corrispondenti ad un massimo locale nel livello di modificazione di una qualche rappresentazione multi-dimensionale del segnale vocale. Nel caso in esame la rappresentazione del segnale è data dalle 80 uscite del MSUP: 40 per il modulo rivelatore di inviluppo e 40 per quello rivelatore di sincronia. In una seconda fase si procede, per associazioni successive, a "clusterizzare" localmente porzioni simili di segnale, ripetendo la procedura fino a che l'intero stimolo viene rappresentato da un unico evento. In pratica si paragonano regioni adiacenti del segnale e ad ogni passo della procedura si fondono in un unica regione quelle due regioni adiacenti che risultano essere maggiormente associate fra loro sulla base di un determinato criterio di somiglianza. Tenendo traccia della distanza a cui due regioni vengono fuse assieme, la descrizione multi-livello può essere visualizzata in forma di "dendrogramma", come evidenziato in Fig.5. Sovrapponendo a questa struttura una linea orizzontale in differenti posizioni verticali, varie ipotesi di segmentazione possono essere effettuate ed analizzate da esperti fonetisti al fine di estrarre l'allineamento più opportuno.

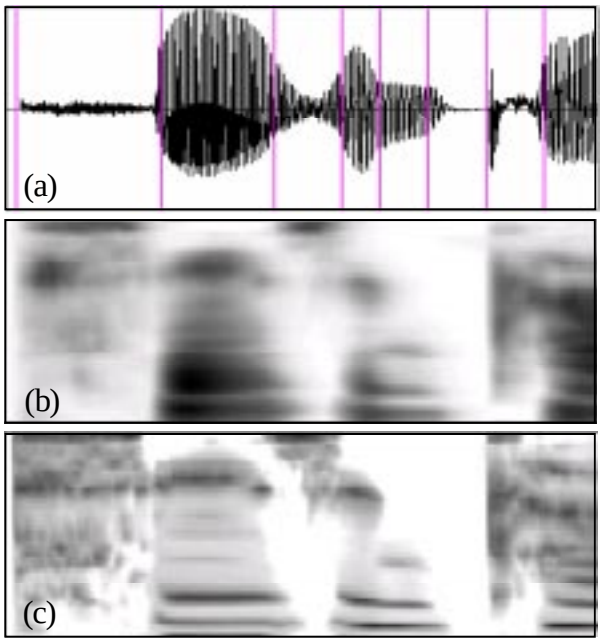

Figura 3. Applicazione del MSUP alla frase "Susan ca(n't)": forma d'onda e segmentazione manuale (a), parametri relativi al rivelatore di inviluppo (b) e di sincronia (c).

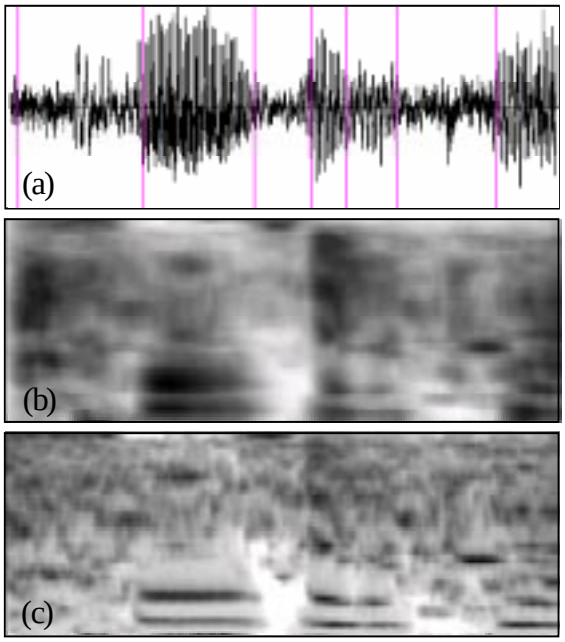

Figura 4. Applicazione del MSUP alla stessa frase di Fig. 3 acquisita in condizioni "rumorose": forma d'onda (a) e parametri di uscita del modello (b), (c).

Considerando la segmentazione di grosse basi dati vocali una tale procedura consente, da una parte, di velocizzare enormemente i tempi di esecuzione, in quanto gli istanti di segmentazione sono posizionati automaticamente dal sistema senza alcun intervento manuale il quale richiederebbe ovviamente di avere accesso ad informazioni visive ed uditive relative al segnale vocale in analisi e, dall'altra, di rendere coerenti gli errori di segmentazione in quanto, se presenti, risultano essere infatti facilmente identificabili e categorizzabili in seguito alla natura algoritmica della procedura.

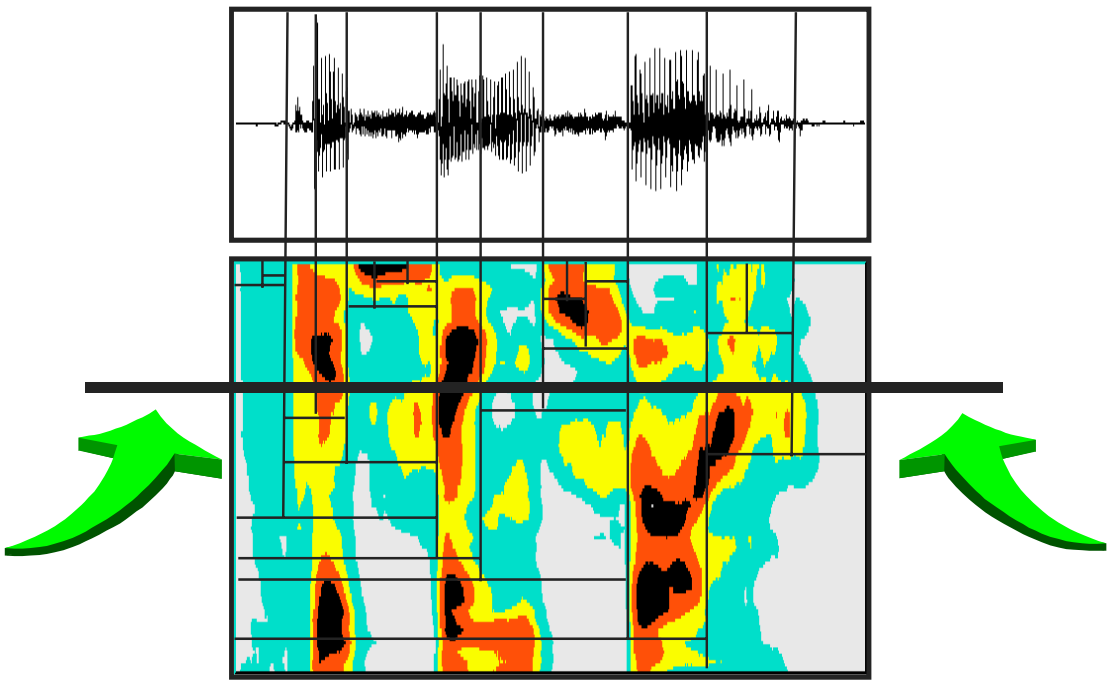

Figura 5. Visualizzazione della procedura di segmentazione semi-automatica sulla base del "dendrogramma" prodotto dal sistema.

In Fig.6 sono illustrati dendrogramma e segmentazione risultante relativi all'analisi della frase "...Susan ca(n't)..." (vedi Fig. 3) mediante MSUP. Fig. 7 si riferisce sempre alla stessa frase considerando però in ingresso i parametri derivanti da una tipica analisi FFT. Come evidente da un confronto di Fig. 6a con Fig. 7a, la struttura multi-livello costruita dall'algoritmo di segmentazione con i parametri relativi al MSUP risulta essere ben più chiara e di ben più facile interpretazione di quella relativa al medesimo algoritmo costruita utilizzando i parametri FFT.
Ad analoghe considerazioni si perviene analizzando Fig. 8 e Fig. 9 che si riferiscono all'analisi della stessa frase registrata però in condizioni di elevato rumore (vedi Fig. 4). Il vantaggio dell'utilizzazione dei parametri relativi al MSUP (Fig. 8a) rispetto a quelli relativi all'analisi FFT (Fig. 9a) risulta essere ancor più evidente in quest'ultimo caso dove infatti l'analisi di Fourier difficilmente riesce a separare il segnale vocale effettivo dal rumore sovrastante a differenza di quanto sembra fare il MSUP. Un analisi quantitativa delle prestazioni del sistema in riferimento ad un esperto è riportata in [4]. Per l'implementazione dell'analisi multi-livello ed anche in funzione dell'efficace interfaccia grafica fornita è stato utilizzato un interessante strumento software, denominato "SONOGRAFIA" e realizzato per Personal Computer in qualità di ausilio alla segmentazione manuale del segnale vocale, nell'ambito del progetto ESPRIT-BRA ACCOR [7]. Per quanto riguarda invece il MSUP, è stata utilizzata una simulazione implementata in FORTRAN presso il Centro di Studio per le Ricerche di Fonetica [3]. Il tempo di calcolo necessario per analizzare il segnale vocale tramite tale modello corrisponde a circa 200 volte il tempo reale (DEC-VAX-Station 3100). Il modello attualmente è da considerarsi assai inefficiente dal punto di vista di una sua ottimizzazione. Una possibile parallelizzazione del modello, consentita dalla sua particolare struttura, non è stata infatti ancora esplorata. Attualmente la stessa simulazione è stata implementata in un DSP "floating point" [8] con tempi di calcolo di circa 10 volte il tempo reale.

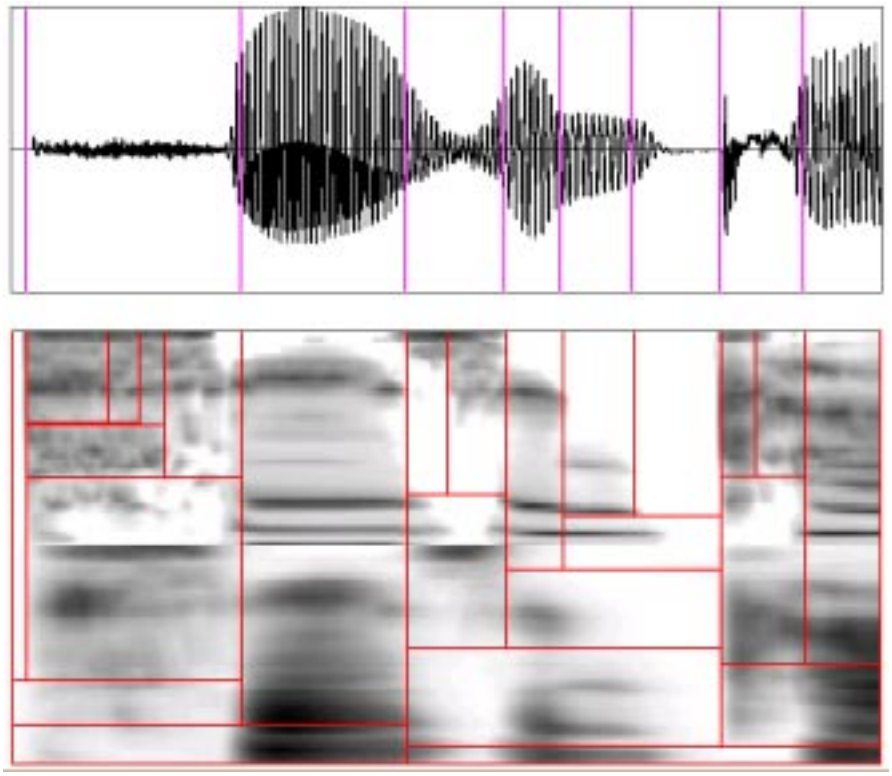

Figura 6. Dendrogramma e segmentazione risultante (MSUP) per la frase "Susan ca(n't)" (vedi Fig. 3).

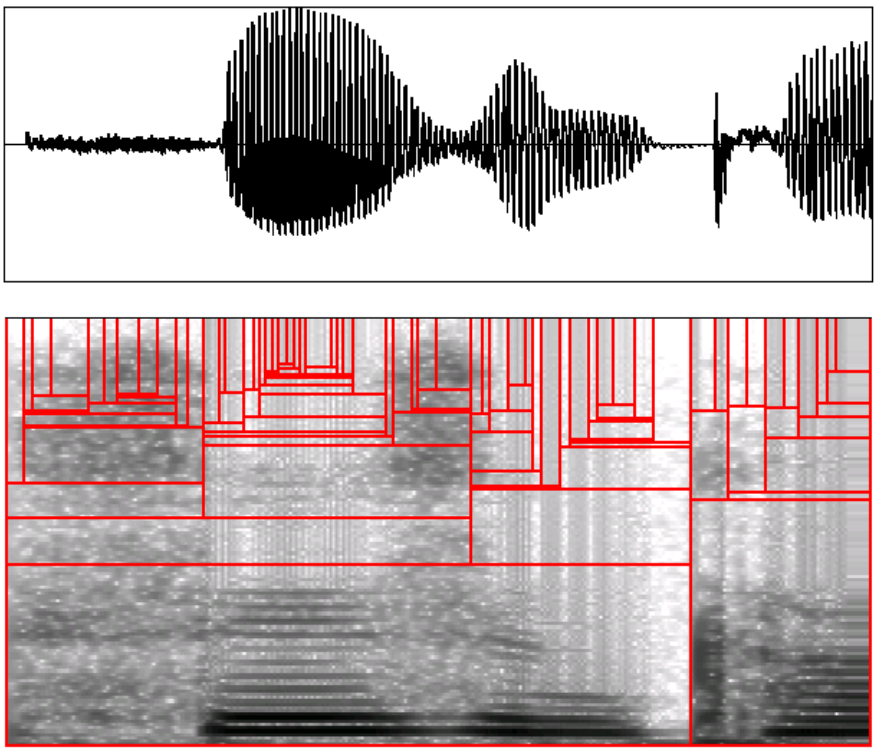

Figura 7. Dendrogramma e segmentazione risultante (FFT) per la frase "Susan ca(n't)" (vedi Fig. 3).

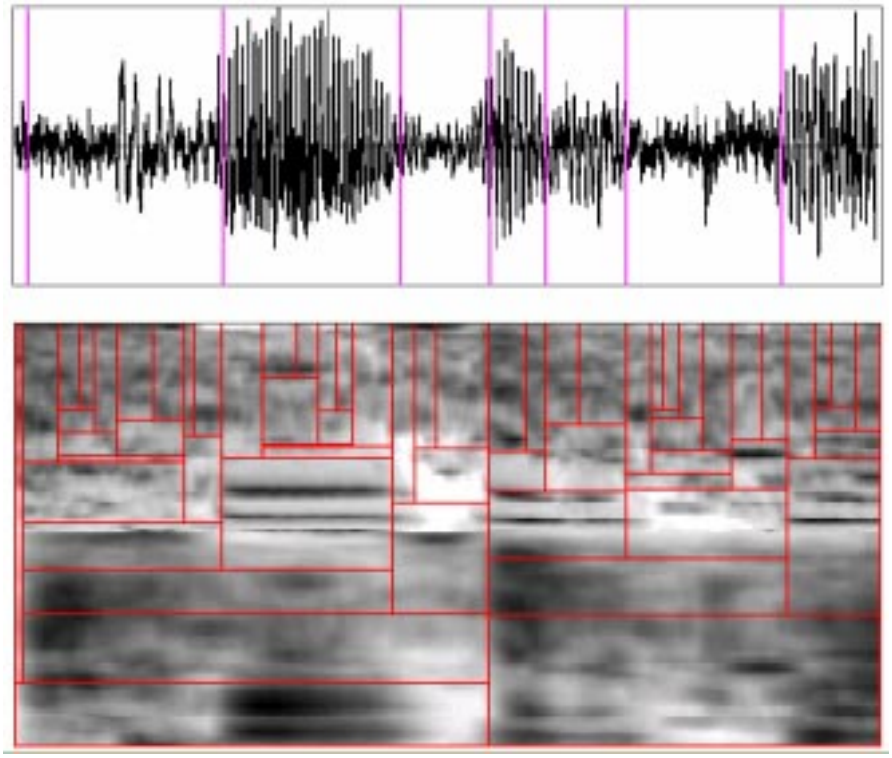

Figura 8. Dendrogramma (MSUP) per la frase "Susan ca(n't)" "rumorosa" (vedi Fig. 4).

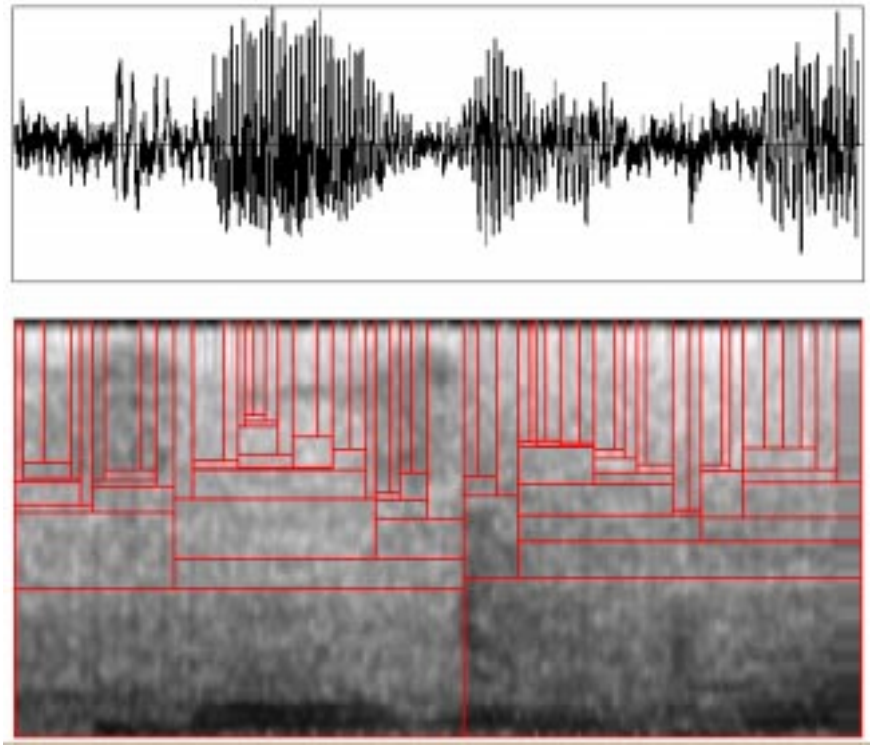

Figura 9. Dendrogramma (FFT) per la frase "Susan ca(n't)" "rumorosa" (vedi Fig. 4).

RISULTATI

Per valutare l'affidabilità del sistema l'intera frase di riferimento ("Che senso ha scaraventare vagonate di giapponesi dentro gli Uffizi. Fargli fare a passo di carica il giro delle sale. Farli uscire, dove trovano soltanto, come documentazione di quello che hanno visto, le schifezze che gli ambulanti, che abusivamente stanno davanti agli Uffizi, gli vendono, a prezzi per altro anche esosi."), utilizzata in tutti i lavori di questo volume, è stata analizzata dal MSUP e successivamente segmentata in modo semi-automatico mediante il sistema appena descritto. La stessa frase inoltre è stata segmentata manualmente seguendo le regole esposte in [9] e sfruttando ausili sia visivi che uditivi (PTS versione 4.40 [10]). In Figg. 10 e 11, rispettivamente per le frasi "Che senso ha" e "gli vendono", sono illustrati: l'uscita del MSUP, la corrispondente segmentazione semi-automatica prodotta dal sistema. In Fig 12 è invece illustrato un'istogramma delle discrepanze fra gli istanti di segmentazione posizionati manualmente e quelli posizionati automaticamente (ELSA [11]). Utilizzando differenti criteri di errore, cioè considerando errori di segmentazione quei casi in cui non c'è corrispondenza fra marker manuali e automatici per più di 10,20 o 30 ms, in altre parole, quei casi in cui i marker automatici sono posizionati al di fuori di una finestra di 20,40 o 60 ms centrata sulla posizione dell'istante di segmentazione di riferimento posizionato manualmente, sempre nella Figura 12 sono indicate anche le differenti percentuali di corretta segmentazione. Mantenendo il criterio di errore su +/-20ms, utilizzato in ambito europeo come criterio standard per valutare tali sistemi la percentuale di errore raggiunge circa il 17%, risultato questo che può sicuramente ritenersi di ottimo livello, vista la difficoltà del compito in esame.

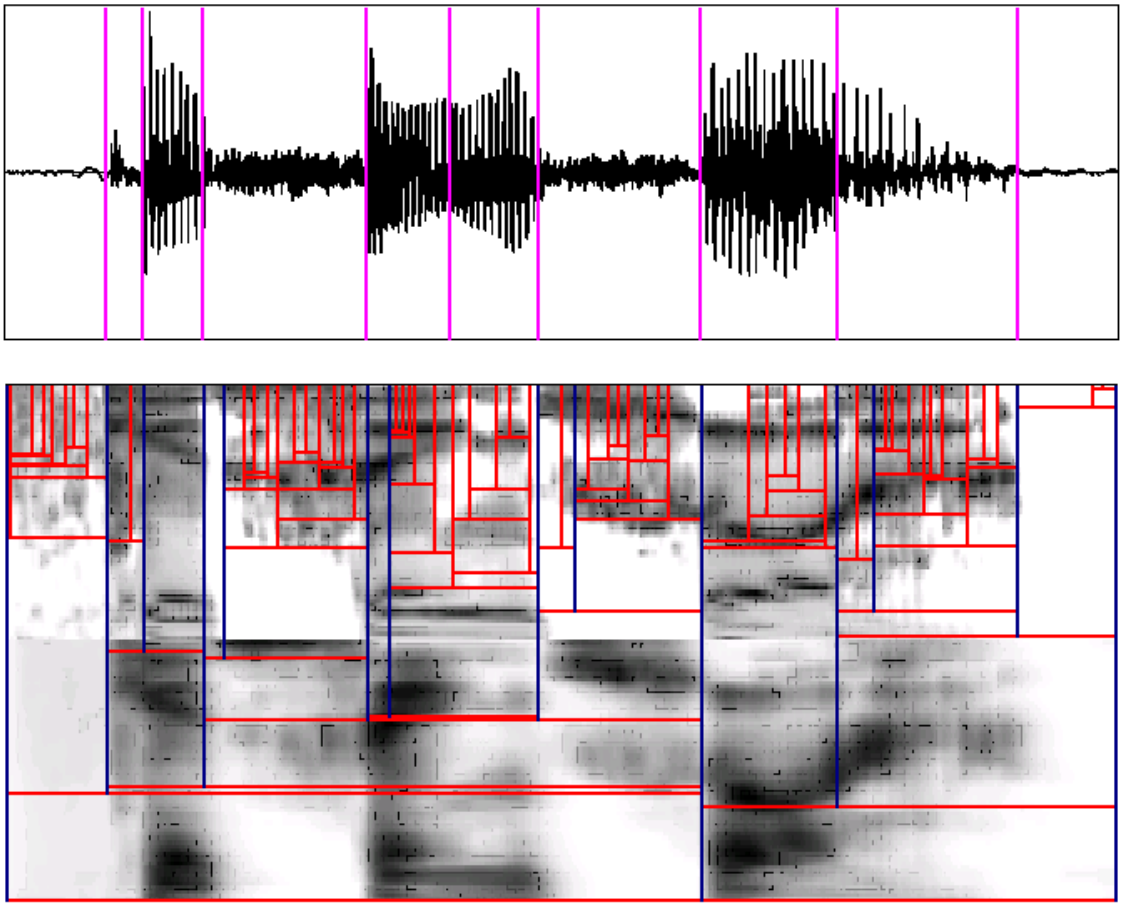

Figura 10. Applicazione del MSUP e dell'algoritmo di segmentazione alla frase "Che senso ha" (vedi testo).

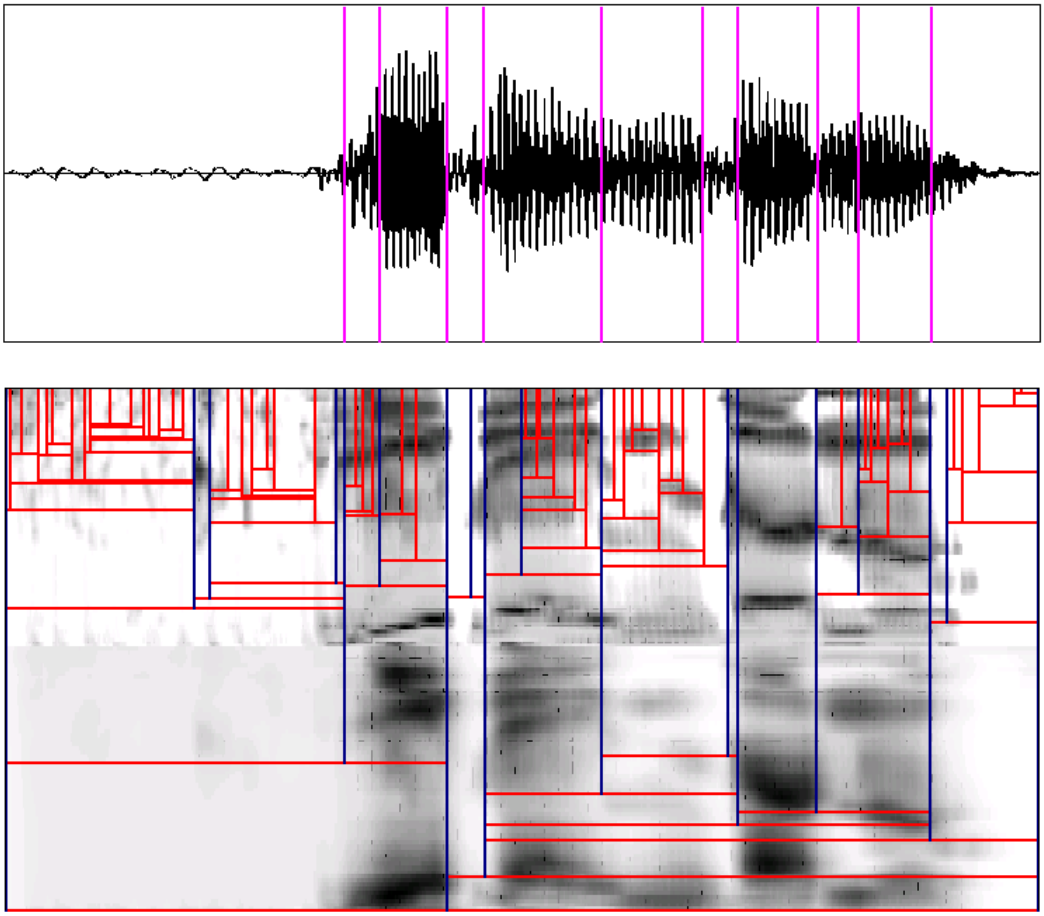

Figura 11. Applicazione del MSUP e dell'algoritmo di segmentazione alla frase "gli vendono" (vedi testo).

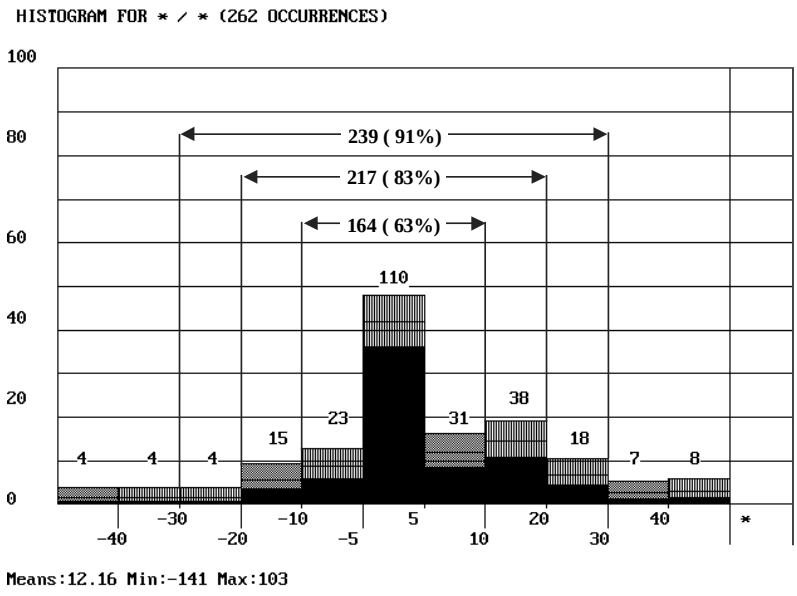

Figura 12. Istogramma degli errori di segmentazione sull'intera frase di riferimento "Che senso ha scaraventare....".

CONCLUSIONI

L'utilizzazione di un modello del sistema uditivo periferico e di un sistema di analisi multi-livello hanno consentito la realizzazione di un efficace sistema semi-automatico di segmentazione del segnale vocale anche in condizioni rumorose. L'utilizzazione di un tale sistema da parte di esperti fonetisti consentirebbe, da una parte, di ridurre enormemente i tempi di segmentazione di grosse basi dati vocali, essendo gli istanti di segmentazione posizionati automaticamente dal sistema senza richiedere l'intervento umano, dall'altra di rendere coerenti i risultati della segmentazione, in quanto, data la natura algoritmica del sistema, gli eventuali errori di segmentazione risulterebbero essere facilmente identificabili e categorizzabili a differenza di quelli umani. Eliminando infine l'intervento umano sulla decisione finale del livello di segmentazione voluto e sostituendolo con un algoritmo di ricerca dell'allineamento ottimo del segnale vocale con la sua corrispondente trascrizione, all'interno della struttura multi-livello, l'algoritmo può diventare completamente automatico, non solo per quanto riguarda la segmentazione o allineamento temporale, ma anche per quanto riguarda il labelling o l'etichettatura del segnale vocale [2]. L'applicazione del sistema alla frase di riferimento ha prodotto dei risultati di segmentazione di sicuro interesse e rilevanza per il proseguimento di questo studio.

BIBLIOGRAFIA

[1] S. Seneff (1988), "*A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing*", Journal of Phonetics, January 1988, pp. 55-76.

[2] V.W. Zue, J. Glass, M. Philips and S. Seneff, "*Acoustic Segmentation and Phonetic Classification in the SUMMIT System*", Proc. IEEE-ICASSP 1989, paper S8.1, pp. 389-392.

[3] P. Cosi, Y. Bengio and R. De Mori,(1990), "*Phonetically-Based Multi-Layered Neural Networks for Vowel Classification*", Speech Comm., Vol. 9, N. 1, Feb 1990, pp. 15-29.

[4] P. Cosi, "*Ear Modelling for Speech Analysis and Recognition*", ESCA Workshop-92, Sheffield, 7-9 Apr, 1992.

[5] M.J. Hunt and C. Lefebvre,"*Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model*", Proc. of ICASSP-88, April 1988, pp.215-218.

[6] J.R. Glass, "*Finding Acoustic Regularities in Speech: Application to Phonetic Recognition*", Ph. D. thesis, Massachusetts Institute of Technology, May 1988.

[7] A. Marzal and J. Puchol, "*Sonografia: an Interactive Segmentation System of Acoustic Signals based on Multilevel Segmentation for a Personal Computer*", ESPRIT II BRA ACCOR Periodic Progress Report 2, 15 Apr 1991, Vol. 3.

[8] P. Cosi, L. Dellana, G.A. Mian and M. Omologo (1991), "*Auditory Model Implementation on a DSP32C-Board*", Proc. GRETSI-91, Juan Les Pins, 16-20 Sep 1991.

[9] P. Cosi, D. Falavigna and M. Omologo, "*A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies*", Proc. of EUROSPEECH-91, Genova, 24-26 Sep 1991, pp. 693-696.

[10] J.C. Caerou. J.M. Dolmazon, A.EL. Badmoussi, K. Jones and B. Barry,"*PTS SOFTWARE v. 4.40: USER MANUAL*", SAM-ESPRIT document.

[11] C. Bourjot, A. Boyer and D. Fohr, "*Semi Automatic Labelling Assessment Software*" SAM-ESPRIT document.